INTERNATIONAL JOURNAL OF
-EXERCISE SCIENCE-

Original Research

Cooper's 12 Minute Run Systematically Underestimates VO_2max in Collegiate Team Sport Athletes: Validity and Classification Accuracy

Edgard MKVK Soares[†1,2#], Robson Silva[*2,3#], Carlos JG da Cruz[†2,4], Luiz GG Porto[‡2], Keila E Fontana[‡2], Guilherme Eckhardt Molina[‡2], Lucia Kobayashi[†2]

[1]Department of Health and Human Physiological Sciences, Skidmore College, Saratoga Springs, NY, USA; [2]Exercise Physiology Laboratory, University of Brasilia, Brasilia-DF, Brazil; [3]Graduate Program of Physical Education, Catholic University of Brasilia, Brasilia-DF, Brazil; [4]Cardiac Autonomic Function Research Group, Centro Universitário Euro Americano-UNIEURO, Brasilia-DF, Brazil

[†]Denotes early-career Investigator, [*]Denotes student investigator, [#]Denotes equal contribution, ‡Denotes established investigator

Abstract

International Journal of Exercise Science 19(2): 2012, 2026. The Cooper 12-Minute Run Test (CT12) is widely used to estimate cardiorespiratory fitness (CRF), yet its validity in collegiate team-sport athletes remains uncertain. This study aimed to compare the CT12-estimated VO_2max with laboratory-measured VO_2max and evaluate CT12's accuracy in classifying CRF levels. Seventy-three varsity athletes (38% women; age 21.6 ± 2.8 y; BMI 23.5 ± 3.3 kg/m²) completed CT12 on a 400-m track and a ramp-incremental treadmill test with breath-by-breath gas analysis, 2–7 days apart in randomized order. VO_2max was classified by age/sex criteria. Agreement statistics included paired tests, Pearson correlation, Bland–Altman bias, and 95% limits of agreement (LOA), 95% confidence intervals (CI), and diagnostic indices (sensitivity, specificity, positive/negative predictive value) for detecting good, excellent, and superior CRF. Field (soccer/rugby) and court (futsal/handball) subgroups were analyzed. CT12 on average underestimated VO_2max by 6.0 $mL(kg.min)^{-1}$ ($p<0.001$; $r=0.69$), with wide LOA: −19.8 to +7.9 $mL(kg.min)^{-1}$. CT12's sensitivity to identify good or better CRF was low (58% [95% CI 45-69%]). However, it demonstrated superior specificity (86% [95% CI 60-96%]) and exhibited an excellent positive predictive value of 94% (95% CI 82-99%). Similar patterns were observed for classifications of excellent and superior CRF. We conclude that in collegiate team-sport athletes, CT12 underestimates VO_2max with substantial individual error. Nevertheless, a positive CT12 classification (good/excellent/superior) is trustworthy due to high specificity/PPV, whereas a negative result warrants further evaluation. These findings endorse CT12 as a practical screening instrument for elevated CRF levels, rather than a replacement for laboratory VO_2max when precise measurement is essential.

Keywords: Bland-Altman, aerobic power, agreement analysis, screening, field test

Introduction

Cardiorespiratory fitness is a physical fitness component and crucial for elite athletic performance[1–4] and is generally expressed as maximal oxygen uptake (VO_2max). VO_2max is the maximal rate at which oxygen is consumed by active muscle during exercise and reflects the efficiency of the integration between the respiratory, cardiovascular, and neuromuscular systems.[5–7] VO_2max plays a vital role in exercise physiology, with its evaluation being essential for coaches and trainers,[3,8] in exercise prescription,[6] physical fitness assessment,[4,9] and monitoring.[10]

*Corresponding Author: Edgard MKVK Soares; Email: esoares@skidmore.edu

Unfortunately, direct measurement of VO_2max using gas analysis is costly, requires a specialized workforce, a carefully controlled laboratory setting, and is time-consuming.[7,11] Therefore, this procedure may be logistically or financially impractical in numerous situations,[11–13] thereby presenting barriers to its implementation in various middle and low-income countries where the gold-standard method is not widely available, primarily due to associated costs.

In addition, among collegiate athletes, it is challenging to reconcile the demanding schedule of a student-athlete with laboratory availability, which might be utilized primarily for research purposes in environments with more constrained budgets. Therefore, conducting multiple VO_2max assessments within a team-sports environment poses significant challenges concerning time management and financial resources. The treadmill exercise test with gas analysis, including necessary post-test procedures, generally requires approximately 60 minutes or more to complete. This constraint may impede the efficient evaluation of physical fitness when multiple athletes need to be tested within a limited timeframe. Thus, valid and reliable indirect methods to estimate VO_2max are warranted. Indirect tests should be affordable and fast, ideally allowing the evaluation of several people in a single day or simultaneously.[7,12]

Various methods have been proposed to estimate VO_2max, including questionnaires,[13] submaximal and maximal exercise tests, which may be performed walking, running, in a track-and-field setting, on a treadmill, or even biking.[11,12] Often, tests use a time-trial result or the distance covered in an allotted time to estimate VO_2max. In a recent meta-analysis, Cooper's 12-Minute Run Test (CT12) was considered one of the best options among the indirect tests.[12] Few studies have examined the agreement between the CT12 and the gold standard measurement in adults. However, findings are usually limited to active men and women[7,14,15] or recreational runners[16] only. Sometimes results suffer from methodological limitations, e.g., comparing CT12 to the VO_2max obtained from a bicycle ergometer test rather than a running one[17]. We are not aware of such validation studies in collegiate athletes, especially from team sports.

Indirect tests may also be used to evaluate cardiorespiratory fitness (CRF) level, i.e., to classify one's CRF as adequate or inadequate using the VO_2max estimated by the test. Several organizations adopted the CT12 to evaluate CRF. The Brazilian Army, Air Force, and Navy all require a minimum CRF level before admission, which is evaluated using the CT12. The Brazilian Federal District Military Firefighter Brigade annually evaluates the CRF of the entire Brigade using the CT12. More than 4000 firefighters are assessed within 2-3 months.[18] Other organizations that use CT12 are the Fédération Internationale de Football Association (FIFA) and the Union of European Football Associations (UEFA), both of which require their soccer referees to maintain a minimum CRF level to obtain and maintain certification as a FIFA or UEFA referee. Bartha et al[19] investigated FIFA-qualified referees and all Hungarian soccer referees. They observed that the higher the referee's qualification, the better he performed in CT12, with significant differences across all categories, e.g., FIFA-qualified, Hungarian-elite-qualified, and Hungarian-county-level-qualified. Despite this result, it remains unclear whether the CT12 accurately classifies CRF levels in other scenarios. To the best of our knowledge, no study has yet investigated CT12's accuracy in discriminating between higher and lower CRF levels in team-sport athletes.

Despite each sport modality having its particularities and needs, if CT12 is shown to be valid on college-level athletes from different team sports, CT12 can be an attractive, low-cost, and practical tool for athletic trainers to classify CRF levels. Therefore, this study aimed to evaluate the validity of VO_2max estimates from CT12 using the gold-standard gas-analysis maximal ramp-incremental treadmill test in collegiate athletes. We also aimed to compare the accuracy of CRF levels derived from the estimated VO_2max from the CT12.

Methods

Participants

A sample size analysis was conducted using G*Power 3.1, which indicated that 68 participants were required (effect size = 0.7, power = 0.80, significance = 0.05, two-tailed). This effect size was hypothesized to represent a difference of 3.5 mL(kg.min)$^{-1}$ in two groups with a coefficient variation of approximately 10% (e.g., group average VO_2max of 50 mL(kg.min)$^{-1}$ and SD of 5 mL(kg.min)$^{-1}$). We recruited slightly more participants to account for potential sample loss or dropouts (≈5%). This study involved collegiate athletes (n=73) who were members of the University's varsity teams and had been engaged in training and competition in varsity sports for at least six months across various modalities: soccer (18 males), rugby (6 males and 11 females), futsal (12 females), and handball (21 males and 5 females). Sex was self-identified by athletes. To respect the similarities and differences between modalities, we also analyzed them as two sub-groups, the field sports group (FG) - soccer and rugby - and the court sports group (CG) - futsal and handball. All volunteers provided written informed consent. The Human Research Ethics Committee of the School of Health Sciences (CEP-FS) of the University of Brasilia approved this study (technical statement number 2,445,992). This research was carried out entirely in accordance with the ethical standards of the *International Journal of Exercise Science.*[20]

Protocol

A cross-sectional study was conducted to evaluate the validity of CT12 in collegiate athletes using convenience, non-probabilistic sampling. Participants performed the CT12 and the maximal ramp-incremental treadmill test (TT) in a randomized order, with 2 to 7 days between them. Participants were instructed to sleep from six to eight hours before test days and to refrain from exercising at least 24 hours before tests; alcohol and stimulating substances such as caffeine should be avoided for at least 48 hours. Trials happened between 8-10 AM or 4-6 PM, according to the athlete's availability. All sessions were conducted during the dry season of the city of Brasilia; thus, no rain or light wind was present on the testing days. Also, testing hours were chosen to avoid higher temperatures and lower humidity. All tests were performed at the same hour of the day for each subject. Participants would first go to the Exercise Physiology Laboratory to meet the researchers, perform anthropometric measurements (first session only), and then, randomly, proceed to one of the tests.

For the Cooper's 12-Minute Run Test (CT12*)***,** participants were instructed to complete as many laps as possible on an outdoor 400-meter track-and-field track during the 12-minute test duration. Participants performed their tests one at a time to minimize any potential bias of an athlete trying to catch up to a faster athlete or slowing down to be closer to a slower athlete. A member of the research team would blow a whistle to signal the end of the test, and participants would immediately stop running. The total distance covered was measured with 5-meter precision, and used to estimate VO_2max in mL(kg.min)$^{-1}$ using Cooper's equation (VO_2maxCT): VO_2max = distance covered in meters – 504.9/44.73.[21]

For the Ramp-Incremental Treadmill Test (TT)**,** oxygen uptake (VO_2) was measured using a breath-by-breath gas analysis system (Metalyzer 3B, Cortex, Leipzig, Germany), which was calibrated and used according to the manufacturer's recommendations. HR was measured using a 3-lead electrocardiogram (Micromed, Brazil). Before performing a maximal ramp-incremental treadmill test (TT), participants would sit for 3 minutes, then stand for 1 minute, and perform a 3-minute warm-up at 3.0 km/h with 0% incline. The ramp-incremental treadmill test would then begin at 3.0-4.0 km/h with a constant 3.0% grade throughout the test. Speed was set to increase at a rate of 0.70-0.90 km/h per minute to ensure a test duration of 8-12 minutes.[22,23] After reaching maximal exertion (voluntary fatigue), participants performed a five-minute active recovery (2.4 km/h and 2.5% incline). At least three of the following four criteria had to be met to ensure maximal exertion: respiratory quotient ≥ 1.1, Borg rating perceived exertion > 17 (6 to 20 points scale), no increase in HR after an increase in workload, and a visually observed VO_2 plateau despite an increase in workload.[11] An experienced evaluator confirmed the presence of VO_2 plateau (VO_2max) in at least two of the final loads of the test through visual inspection of the VO_2 per workload plot.[22]

To assess Cardiorespiratory Fitness (CRF) levels, Cooper's classification criteria were utilized to categorize CRF based on age and sex.[24] We did not utilize sport-specific CRF classification criteria given that the CT12 is not tailored specifically for certain sports, and our objective was to evaluate its applicability in assessing large cohorts of athletes from diverse modalities and backgrounds. As a "screening tool," the CT12 may prove useful in identifying individuals requiring sport-specific assessment, wherein such criteria would subsequently be applied. The total distance covered corresponding to the classifications: very weak, weak, regular, good, excellent, and superior was transformed into $mL(kg.min)^{-1}$ using Cooper's formula and used to evaluate CRF levels in both tests.

Statistical Analysis

Data were reported as mean ± SD. Normality was implied, considering the central limit theorem.[25] The t-test was used to compare means due to its robustness in different types of distribution.[26] Statistical significance was considered at $p<0.05$ for all analyses. Estimated VO_2max (VO_2maxCT) was compared to the gold standard (VO_2maxTT) using a paired Student T-test. Agreement between measurements was evaluated using the Bland-Altman analysis plot, Pearson's correlation coefficient, and a one-sample t-test, the latter to assess whether the bias (the difference between VO_2maxCT and VO_2maxTT) was significantly different from zero. Bland and Altman's analysis plot has been described thoroughly elsewhere.[27,28]

For the subgroup analysis (FG vs. CG), an independent T-test was performed to compare between-groups variables. The proportion of women in both subgroups was compared using the Chi-square test. Agreement between VO_2maxTT and VO_2maxCT was analyzed independently in the FG and CG using the same methods described above.

Strength of the Pearson's correlation coefficient was interpreted as small ($r \geq 0.1$ and < 0.3), moderate ($r \geq 0.3$ and < 0.5), large ($r \geq 0.5$ and < 0.7), very large ($r \geq 0.7$ and < 0.9), and extremely large ($r \geq 0.9$). Effect sizes of differences were analyzed using Cohen's "*d*" and interpreted as small ($d \geq 0.2$ and < 0.5), moderate ($d \geq 0.5$ and <0.8), and large ($d \geq 0.8$).[29,30]

We also performed three agreement analyses to compare differences between CRF classifications resulting from CT12 and TT test results. First, the individual VO_2maxTT and VO_2maxCT results were classified using the previously explained criteria,[24] with VO_2maxTT considered the referent method in a contingency table analysis. The three analyses had the following outcomes of interest: (1) athletes who possess a good CRF (classified as good, excellent, or superior CRF); (2) athletes who have an excellent CRF (classified as excellent or superior CRF); (3) athletes who have a superior CRF (classified as superior CRF). In the contingency table, true positives (TP) were those that were classified by both tests (TT and CT12) as having the outcome of interest. True negatives (TN) were those that were classified by both tests as not having the outcome of interest. False positives (FPs) were those classified only by CT12 as having the outcome of interest. False negatives (FN) were those that were classified only by the TT as having the outcome of interest.

The following agreement indices were used to evaluate CT12's accuracy as a tool to detect good, excellent and superior CRF level: 1) total agreement, the sum of the percentage of TP and TN; 2) sensitivity, the ratio between TP and TP+FN; 3) specificity, the ratio between TN and TN+FP; 4) positive predictive value, the ratio between TP and TP+FP; 5) negative predictive value, the ratio between TN and TN+FN. All epidemiological indices were calculated as point estimates and 95% confidence intervals (CIs).

Results

Table 1 describes the volunteers' characteristics. The similarity of anthropometric variables, sex proportions, and age between field and court athletes allowed us to analyze them independently (FG and CG) and as a single group (Table 1).

Table 1. Participant's descriptive statistics.

	All athletes (n = 73)	FG (n = 35)	CG (n = 38)	p^a	d^a
Female athletes (%)	38.4	31.4	44.7	0.24^b	-
Age (years)	21.6 ± 2.8	21.9 ± 3.0	21.3 ± 2.6	0.36	0.21
Height (cm)	172.3 ± 9.3	172.8 ± 8.8	171.8 ± 9.9	0.66	0.11
Body mass (kg)	70.1 ± 12.5	68.9 ± 10.6	71.2 ± 14.2	0.45	0.18
BMI (kg/m^2)	23.5 ± 3.3	23.1 ± 3.3	23.9 ± 3.3	0.27	0.24
BF%	17.7 ± 10.0	16.4 ± 11.2	19.0 ± 8.6	0.27	0.26

Values are mean ± SD unless otherwise stated. FG: field sports group; CG: court sports group; BMI: body mass index, BF%: body fat percentage; [a]: Field vs. Court group, d: Effect Size of FG vs. CG (Cohen's "d"); [b]: Only the percentage of females is shown; [c]: Chi-square test.

As seen in Table 2, FG and CG had similar VO_2max when measured by the gold-standard method, the maximal ramp-incremental treadmill test with gas analysis (VO_2maxTT), p=0.19, *d*=0.32. However, they presented a significantly different estimated VO_2max (VO_2maxCT) – p=0.02, *d*=0.58. VO_2maxTT and VO_2maxCT were significantly different when compared in the entire group as well as in the subgroups (p<0.0001, *d*=0.8–1.2) – Table 2. Bland and Altman's plot analyses indicated that CT12 is inaccurate for both the whole group and the analyzed subgroups, as shown by significant bias and wide limits of agreement (Figure 1).

Table 2. VO_2maxTT and VO_2maxCT (mean ± SD) in the main group, field, and court subgroups and agreement indices.

Groups	VO_2maxTT mL(kg.min)-1	VO_2maxCT $mL(kg.min)^{-1}$	Bias $mL(kg.min)^{-1}$	Bias 95%CI $mL(kg.min)^{-1}$	*d*	r
All athletes (n=73)	44.9 ± 7.8	38.9 ± 9.5*	$-6.0^{\#}$ ± 6.9	4.4-7.6	0.81	$0.69^{\dagger}$
Field athletes (n=35)	46.2 ± 9.7	41.7 ± 9.7*	$-4.5^{\#}$ ± 5.9	2.5-6.5	0.99	$0.82^{\dagger}$
Court athletes (n=38)	43.8 ± 5.4	36.4 ± 8.5*	$-7.4^{\#}$ ± 7.6	4.9-9.9	1.22	$0.49^{\dagger}$

VO_2maxTT: Maximal oxygen uptake measured during the ramp-incremental treadmill test; VO_2maxCT: Maximum oxygen uptake estimated by the Cooper's test; SD = standard deviation; CI = Confidence interval; d: Effect Size (Cohen's "d") of VO_2maxTT vs VO_2maxCT; * = p < 0.0001 For comparison between VO_2maxTT and VO_2maxCT; # = statistically different compared to zero (p<0.0001); r = Pearson's correlation coefficient; † = Significant correlation (p < 0.01).

Secondly, athletes' cardiorespiratory fitness (CRF) levels were evaluated by classifying the VO_2max obtained from both tests (Table 3). Afterward, epidemiological indices were used to assess the agreement between the CRF level provided by the CT12 and the TT results (Table 4). CT12 showed low sensitivity (26-58%), high specificity (86-98%), and excellent positive predictive value (86-94%) for detecting a good, excellent, or superior CRF. These findings were similar when subgroups were analyzed separately (data not shown).

Table 3. Absolute and relative frequency of cardiorespiratory fitness classification according to the test performed (n=73).

Test	Very weak	Weak	Regular	Good	Excellent	Superior
TT	0 (0.0%)	2 (2.7%)	12 (16.4%)	20 (27.4%)	16 (21.9%)	23 (31.5%)
CT12	9 (12.3%)	10 (13.7%)	18 (24.7%)	21 (28.8%)	8 (11.0%)	7 (9.6%)

Values are n(%); TT: ramp-incremental treadmill test, CT12: Cooper's 12-Minute Run/Walk Test.

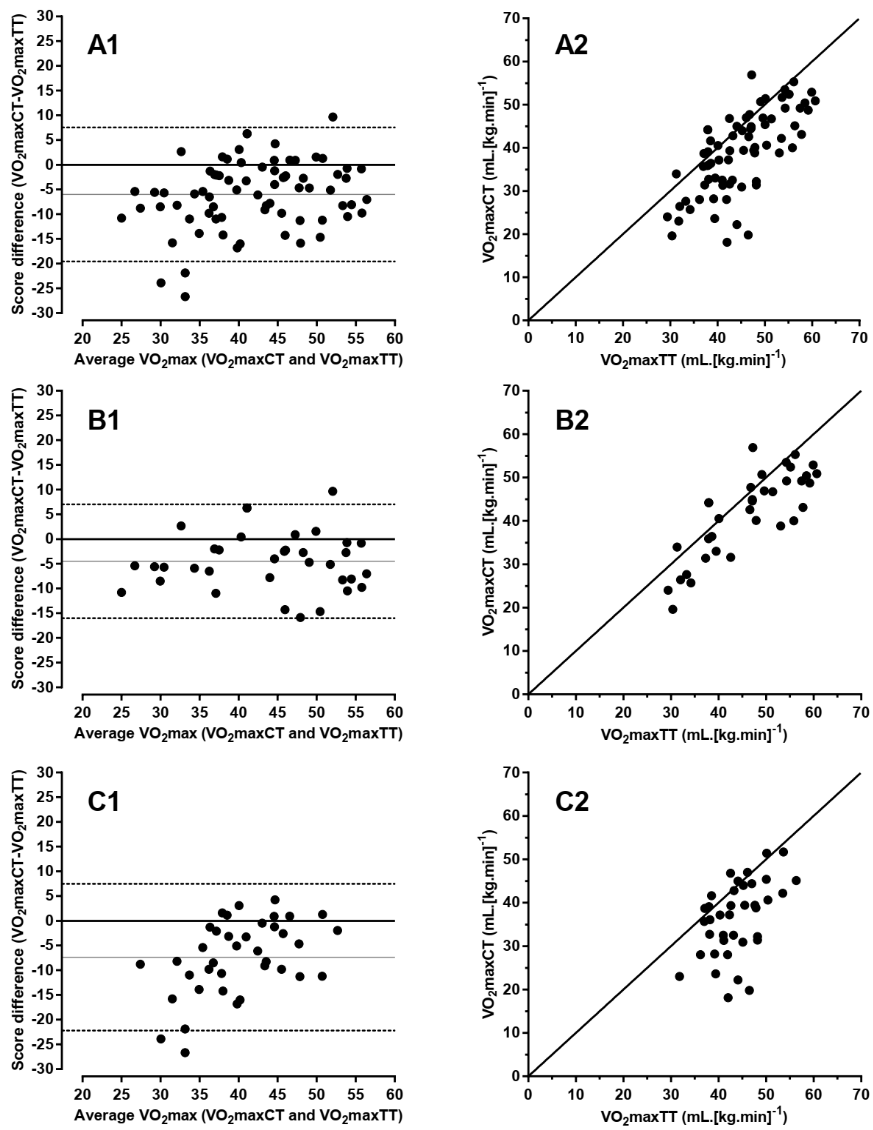

Figure 1. Bland-Altman plot for the whole group (A1) and the two subgroups: field (B1) and court (C1) athletes. Dashed lines are upper and lower agreement limits; the gray line is the mean bias of Cooper's test estimated VO_2max (VO_2maxCT) compared to the directly measured VO_2max from the treadmill test (VO_2maxTT). VO_2maxCT was plotted against VO_2maxTT, together with an identity line, for the entire group (A2) and for the subgroups: field (B2) and court (C2) athletes.

Table 4. Epidemiological agreement indices of CT12's accuracy in detecting good, excellent, and superior cardiorespiratory fitness (n = 73).

			TT (CRF = Good) Yes	No	*Total*	
			n (%)	**n (%)**	**n**	
CT12 (CRF = Good)	**Yes**	**n (%)**	**34 (46.6%)**	**2 (2.7%)**	*36*	***Accuracy: 63.0% (95% CI 51.5 – 73.2)*** ***Sensitivity: 57.6% (95% CI 44.9 – 69.4)*** ***Specificity: 85.7% (95% CI 60.1 – 96.0)*** ***Positive Predictive Value: 94.4% (95% CI 81.9-98.5)*** ***Negative Predictive Value: 32.4% (95% CI 19.6-48.5)***
	No	**n (%)**	**25 (34.2%)**	**12 (16.4%)**	*37*	
	Total	***n***	*59*	*14*	*73*	

			TT (CRF = Excellent) Yes	No	*Total*	
			N (%)	**N (%)**	***N***	
CT12 (CRF = Excellent)	**Yes**	**N (%)**	**14 (19.2%)**	**1 (1.4%)**	*15*	***Accuracy: 64.4% (95% CI 52.9 – 74.4)*** ***Sensitivity: 35.9% (95% CI 22.7 – 51.6)*** ***Specificity: 97.1% (95% CI 85.1 – 99.5)*** ***Positive Predictive Value: 93.3% (95% CI 70.2-98.8)*** ***Negative Predictive Value: 56.9% (95% CI 44.1-68.8)***
	No	**N (%)**	**25 (34.2%)**	**33 (45.2%)**	*58*	
	Total	***N***	*39*	*34*	*73*	

			TT (CRF = Superior) Yes	No	*Total*	
			N (%)	**N (%)**	***N***	
CT12 (CRF = Superior)	**Yes**	**N (%)**	**6 (8.2%)**	**1 (1.4%)**	*7*	***Total Accuracy: 75.3% (95% CI 64.4 – 83.8)*** ***Sensitivity: 26.1% (95% CI 12.5 – 46.5)*** ***Specificity: 98.0% (95% CI 89.5 – 99.6)*** ***Positive Predictive Value: 85.7% (95% CI 48.7-97.4)*** ***Negative Predictive Value: 74.2% (95% CI 62.6-83.3)***
	No	**N (%)**	**17 (23.3%)**	**49 (67.1%)**	*66*	
	Total	***N***	*23*	*50*	*73*	

TT: ramp-incremental treadmill test; CT12: Cooper's 12-Minute Run/Walk Test; CRF: Cardiorespiratory fitness; CI: Confidence interval.

Discussion

Our study aimed to evaluate the validity of VO_2max estimates from CT12 and the accuracy of CRF level classification using CT12-estimated VO_2max. In this study, Cooper's 12-Minute Run Test (CT12) results - VO_2maxCT - significantly underestimated the VO_2max in collegiate athletes compared to the measured VO_2max obtained by the gold standard method (VO_2maxTT). Analysis of field and court athletes separately resulted in similar findings (Table 2 and Figure 1). Bland and Altman's analysis plot showed no proportional bias with wide limits of agreement (Figure A1, B1, C1). VO_2maxCT's mean bias was -6.0 mL(kg.min)$^{-1}$ compared to the VO_2maxTT. The underestimation trend is evident in Figure A2, which shows a large number of participants below the identity line compared to those above it. The same phenomena were observed when the analysis was stratified by field and court subgroups (Figure B2, C2). CRF plays a vital role in high-level performance in team sports.[4,31] Thus, an underestimation of this magnitude may affect coaches' or trainers' decision-making process on an athlete's level of fitness and capacity to play, since VO_2max is associated with the total distance covered during match play in some team sports.[2–4,8] Likewise, VO_2max differed between athletes selected by a professional rugby league and those not selected.[32] The wide limits of agreement raise more concern than the underestimation, because they indicate significant variability in the bias.

When using VO_2maxCT to evaluate CRF level, CT12's underestimation trend can be identified by the higher frequency of participants classified as having a very weak and weak CRF, and fewer participants classified as having a superior CRF compared to the TT results (Table 3). Regarding CT12's capability to detect a good CRF, we observed low sensitivity (58%) but excellent specificity (86%), indicating that CT12 was highly accurate in identifying individuals whose actual CRF (VO_2maxTT) was classified as good. We also highlight its excellent positive predictive value (94%), indicating that athletes who presented good CRF according to CT12 also had good CRF in the TT. Similar behavior was seen using higher cut-points, i.e., excellent and superior CRF (Table 4), and in subgroups (data not shown).

The correlations between VO_2maxTT and VO_2maxCT (Table 2) can be considered large to very large across groups.[29,30] Similar values have been reported in young adults, with most studies showing correlations of: 0.61,[14] 0.65,[16] 0.84,[15] 0.87,[7] 0.92.[33] However, correlation alone is insufficient to assess validity because it reflects the strength of a linear relationship, not whether two methods provide similar values.[28,34] For method comparison, i.e., an agreement analysis, the between-method bias and the distribution of individual errors are of greater informational value. The practical utility of a test should be evaluated on a clinical or functional basis rather than solely through isolated statistical analyses.[27,28]

In the present study, large correlations coexisted with poor agreement (Figure and Table 2). CT12 significantly underestimated VO_2max, and the limits of agreement indicate that, in 95% of cases, CT12 may overestimate the true VO_2max up to 7.9 mL(kg. min)$^{-1}$ or underestimate it by up to 19.8 mL(kg.min)$^{-1}$. This level of error warrants caution when CT12 is used to estimate an individual collegiate athlete's VO_2max. Other studies have observed similar results, with CT12, on average, underestimating VO_2max by 11.9 mL(kg.min)$^{-1}$,[16] 3.8 mL(kg.min)$^{-1}$,[15] and approximately 2.0 mL(kg.min)$^{-1}$, although the latter was not significant (p=0.06).[7] To our knowledge, only one study[14] reported an overestimation (0.4 mL[kg.min]$^{-1}$), but also showed wide limits of agreement (-12.4 to 11.6 mL(kg.min)$^{-1}$ partially resulting from a small sample (n=12). Wide limits of agreement have also been reported by Kravchychyn et al[16] (0.0 to 23.9 mL[kg.min]$^{-1}$), and Penry et al[7] (-10.1 to 6.5 mL[kg.min]$^{-1}$). Penry et al[7] also reported that CT12 exhibited a proportional bias. Although proportional bias may seem plausible when examining the CG subgroup, when observing VO_2maxCT as a function of VO_2maxTT (Figure A2-C2) it is possible to notice that high and low levels of differences (values distant and close to the identity line, respectively) occur at both high and low VO_2maxTT values. Therefore, no proportional bias seems to be present.

It is unknown which specific factors explain the reported underestimation in CT12. Running performance has a strong correlation with the velocity at VO_2max and the velocity at the ventilatory thresholds. Sometimes, these correlations are more robust than the correlation between running performance and VO_2max itself[1,35]. Thus, discrepant CT12 performance in individuals with similar VO_2max may reflect differences in their velocity at VO_2max and ventilatory thresholds, a topic beyond the scope of this study. Considering the nature of their sport, it is also possible that team-sport athletes would perform better on an intermittent test, such as the Yo-Yo Intermittent Recovery Test (YYR)[32]. However, given that the velocity at VO_2max is frequently more strongly correlated with performance on such tests than the VO_2max itself[31], a direct comparison between CT12 and YYR is warranted, which is also beyond the scope of our study.

To the best of our knowledge, this is the first study to perform a comprehensive accuracy analysis of CT12 for detecting good, excellent, and superior CRF (Cooper's CRF classification[24]) in college-level athletes using epidemiological indices. CT12 showed remarkably high positive predictive value, i.e., when someone's CT12 result was classified as a good or better CRF, it was correct in 94.4% of the time. CT12 also showed excellent specificity (85-98%). Thus, CT12 was useful for correctly detecting low CRF in athletes without good, excellent, or superior CRF. As a practical application of these findings, a "positive" result on the CT12 is valuable for coaches and trainers, as athletes performing well on the CT12 have very low odds of having a low CRF level.

Because of the originality of our work, direct comparison with previous CT12 classification studies is not possible. However, other authors have investigated different CRF evaluation methods in other populations. Sartor et al[36] investigated the accuracy of a 45s squat test in detecting good and fair CRF (ACSM criteria). The best mathematical models had a sensitivity of 62-64% and a specificity of 62-63%. Peterman et al[13] investigated 28 non-exercise equations in apparently healthy men and women from the Ball State Adult Fitness Longitudinal Lifestyle Study cohort and found correct classification rates ranging from 34% to 62% (using the Fitness Registry and the Importance of Exercise National Database criteria). In our study, CT12 correctly classified (total accuracy) 63-75% of the athletes, depending on the level of fitness used for comparison. Flouris et al[37] evaluated 70 men using the 20 m shuttle run test to detect low CRF. Using the original VO_2max equation, the authors observed a significant overestimation of CRF, which resulted in very low sensitivity (23%), perfect specificity and positive predictive value (100%), and reasonable negative predictive value (69%). VO_2max overestimation explains these findings; participants with low CRF were classified as having adequate CRF (low sensitivity), but men with adequate CRF were always classified correctly (high specificity). Considering adequate CRF as the outcome of interest, the test would correctly classify 69% of men. In our study, CT12 accurately detected good, excellent, and superior CRF in 86-94% of male and female collegiate athletes.

Despite this usefulness for classification, CT12 should not be used interchangeably with direct VO_2max measurement.[14–16] The underestimation may lead coaches and conditioning specialists to prescribe exercise intensity based on a lower VO_2max, potentially reducing the training load and constraining performance improvements. In practical applications, a training session scheduled to be conducted at a vigorous intensity may, in reality, be performed at a moderate intensity. This discrepancy significantly reduces the intended overload and consequently hampers the anticipated adaptations associated with high-intensity training sessions. This limitation is not unique to the CT12, as other indirect tests also show significant bias,[16,17,38] or wide limits of agreement.[7,14,39,40] We highlight the importance of examining both mean bias and agreement because an indirect test may seem adequate when the average bias is near zero; however, it may be overestimating VO_2max in unfit individuals and underestimating VO_2max in very fit individuals.[7] A perfect indirect test does not exist yet; thus, practitioners should consider reliability,[7,12] time, and available resources when selecting a test to estimate VO_2max.

According to our findings, CT12 is most useful for evaluating CRF when positive results are observed, i.e., when athletes' estimated VO_2max is classified as good, excellent, or superior. In such cases, practitioners can be reasonably confident that the athlete does not have a low CRF level. A negative test result is less informative because CT12 significantly underestimates VO_2max and, consequently, CRF.

Among the limitations of the present study, we highlight the absence of CRF classification criteria specific to the sport modality; however, this limitation is mitigated by the similarity of VO_2max and anthropometric characteristics observed across subgroups and by the absence of an ideal CRF classification for every sport modality. The use of "generic" classification criteria in our study also facilitates the use of CT12 as a screening tool, since sports-specific criteria are often confined to books that are not easily accessible to practitioners. The use of different classification criteria would likely yield different results, although we believe CT12 would retain its core characteristic of average-to-low sensitivity and high specificity. Caution is also necessary when extrapolating our findings to populations that did not participate in this study, such as highly fit college athletes in track and field and combat sports. Future research should compare CT12 with other indirect tests (e.g., Yo-Yo Intermittent Recovery Test), use different CRF classification criteria, and expand the athletic population to include other sports modalities.

Another possible limitation is the athletes' motivation to perform a maximal CT12. It is possible that some athletes did not exert their maximum effort during the CT12, which could have contributed to the underestimation relative to the ramp-incremental treadmill test. To minimize this possibility, the research team supervised all tests, and athletes were adequately instructed to perform to their fullest.[7,12] Notably, our findings are similar to previous reports, with the vast majority of studies also showing VO_2max underestimation by CT12.[14–16] Another possible limitation is the relatively small sample size for an agreement study, i.e., n<100 individuals. However, significant differences were observed in almost all comparisons, making type II error unlikely, and estimated confidence intervals in the Bland-Altman plot likely reflect actual test variability rather than a small sample size. Studies with athletes rarely are able to have large samples, and our sample was relatively large compared to previous studies using smaller groups of physically active people: 12,[14] 15,[16] 21.[7] Lastly, CT12 is known to have elevated reliability; we believe it is unlikely that a second performance would significantly change our results.[7]

In conclusion, Cooper's 12-Minute Run Test (CT12) significantly underestimates VO_2max in collegiate team-sport and shows wide limits of agreement, despite the absence of proportional bias. Nevertheless, the analytical approach used here supports an evidence-based use of CT12 as a screening tool for higher CRF levels in college-level team-sports athletes. A positive CT12 result, that is, when an athlete's CRF is classified as good, excellent, or superior, appears trustworthy and may help coaches and trainers in the decision-making process to assess whether an athlete has an adequate fitness level or requires further evaluation. These findings also support those who already use the CT12, giving confidence that a good or better CRF level obtained from the CT12 is very likely accurate. In contrast, a negative CT12 outcome, i.e., being classified as having a low CRF level, is less useful because of the significant underestimation of the CT12, and some fit athletes may be misclassified as having low CRF and may require further evaluation using other methods.

Although practitioners recognize these limitations, no indirect test is entirely free of bias. CT12 can be applied easily to multiple individuals simultaneously, is low-cost, and does not require a specialized workforce. Thus, it remains a reasonable alternative for coaches and athletic trainers who need an indirect test to evaluate multiple athletes. CT12 may be especially useful when direct CRF assessment is unavailable and may also improve the cost-effectiveness of using the direct testing by reserving it for athletes whose results are doubtful.

Acknowledgements

Author EMKVKS was supported by a scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) finance code 001.

References

1. Alvero-Cruz JR, Carnero EA, Giráldez García MA, et al. Cooper Test Provides Better Half-Marathon Performance Prediction in Recreational Runners Than Laboratory Tests. *Front Physiol*. 2019;10:1349. https://doi.org/10.3389/fphys.2019.01349

2. Helgerud J, Engen LC, Wisloff U, Hoff J. Aerobic endurance training improves soccer performance. *Med Sci Sports Exerc*. 2001;33(11):1925-1931. https://doi.org/10.1097/00005768-200111000-00019

3. Makaje N, Ruangthai R, Arkarapanthu A, Yoopat P. Physiological demands and activity profiles during futsal match play according to competitive level. *J Sports Med Phys Fitness*. 2012;52(4):366-374.

4. Stone NM, Kilding AE. Aerobic conditioning for team sport athletes. *Sports Med Auckl NZ*. 2009;39(8):615-642. https://doi.org/10.2165/00007256-200939080-00002

5. Bertuzzi R, Nascimento EMF, Urso RP, Damasceno M, Lima-Silva AE. Energy system contributions during incremental exercise test. *J Sports Sci Med*. 2013;12(3):454-460.

6. Mann T, Lamberts RP, Lambert MI. Methods of prescribing relative exercise intensity: physiological and practical considerations. *Sports Med Auckl NZ*. 2013;43(7):613-625. https://doi.org/10.1007/s40279-013-0045-x

7. Penry JT, Wilcox AR, Yun J. Validity and reliability analysis of Cooper's 12-minute run and the multistage shuttle run in healthy adults. *J Strength Cond Res*. 2011;25(3):597-605. https://doi.org/10.1519/JSC.0b013e3181cc2423

8. Stølen T, Chamari K, Castagna C, Wisløff U. Physiology of soccer: an update. *Sports Med Auckl NZ*. 2005;35(6):501-536. https://doi.org/10.2165/00007256-200535060-00004

9. Pauw K De, Roelands B, Geus B De, Meeusen R. Guidelines to classify subject groups in sport- science research. *Int J Sports Physiol Perform*. 2013;8:111-122. https://doi.org/10.1123/ijspp.8.2.111

10. Clark NA, Edwards AM, Morton RH, Butterly RJ. Season-to-Season Variations of Physiological Fitness Within a Squad of Professional Male Soccer Players. *J Sports Sci Med*. 2008;7(1):157-165.

11. American College of Sports Medicine. *ACSM's Guidelines for Exercise Testing and Prescription*. 10th ed. (Riebe D, Ehrman JK, Liguori G, Magal M, eds.). Wolters Kluwer; 2018.

12. Mayorga-Vega D, Bocanegra-Parrilla R, Ornelas M, Viciana J. Criterion-Related Validity of the Distance- and Time-Based Walk/Run Field Tests for Estimating Cardiorespiratory Fitness: A Systematic Review and Meta-Analysis. *PloS One*. 2016;11(3):e0151671. https://doi.org/10.1371/journal.pone.0151671

13. Peterman JE, Whaley MH, Harber MP, et al. Comparison of non-exercise cardiorespiratory fitness prediction equations in apparently healthy adults. *Eur J Prev Cardiol*. Published online October 22, 2019:2047487319881242. https://doi.org/10.1177/2047487319881242

14. Batista MB, Cyrino ES, Milanez VF, Silva MJC, Arruda M, Ronque ERV. Maximal oxygen uptake estimate and agreement analysis of the direct measurement with different field tests. *Rev Bras Med Esporte*. 2013;19(6):404-409. https://doi.org/10.1590/S1517-86922013000600005

15. McCutcheon MC, Sticha SA, Giese MD, Nagle FJ. A further analysis of the 12-minute run prediction of maximal aerobic power. *Res Q Exerc Sport*. 1990;61(3):280-283. https://doi.org/10.1080/02701367.1990.10608692

16. Kravchychyn ACP, Alves JCC, Kravchychyn TP, Nogueira GÂ, Machado FA. Comparison between the direct and indirect methods of VO_2max determination in runners. *Rev Bras Med Esporte*. 2015;21(1):17-21. https://doi.org/10.1590/1517-86922015210101412

17. Bandyopadhyay A. Validity of Cooper's 12-minute run test for estimation of maximum oxygen uptake in male university students. *Biol Sport*. 2015;32(1):59-63. https://doi.org/10.5604/20831862.1127283

18. CBMDF. Portaria nº. 29, de 11 de agosto de 2004 – Programa Padrão de Treinamento Físico do Bombeiro Militar nas Unidades da Corporação. Published online 2004.

19. Bartha C, Petridis L, Hamar P, Puhl S, Castagna C. Fitness test results of Hungarian and international-level soccer referees and assistants. *J Strength Cond Res*. 2009;23(1):121-126. https://doi.org/10.1519/JSC.0b013e31818ebb84

20. Navalta JW, Stone WJ, Lyons TS. Ethical Issues Relating to Scientific Discovery in Exercise Science. *Int J Exerc Sci*. 2019;12(1):1-8. https://doi.org/10.70252/EYCD6235

21. Cooper KH. The History of Aerobics (50 Years and Still Counting). *Res Q Exerc Sport*. 2018;89(2):129-134. https://doi.org/10.1080/02701367.2018.1452469

22. Arena R, Myers J, Williams MA, et al. Assessment of functional capacity in clinical and research settings: a scientific statement from the American Heart Association Committee on Exercise, Rehabilitation, and Prevention of the Council on Clinical Cardiology and the Council on Cardiovascular Nursing. *Circulation*. 2007;116(3):329-343. https://doi.org/10.1161/CIRCULATIONAHA.106.184461

23. Myers J, Bellin D. Ramp exercise protocols for clinical and cardiopulmonary exercise testing. *Sports Med Auckl NZ*. 2000;30(1):23-29. https://doi.org/10.2165/00007256-200030010-00003

24. Cooper KH. *Aerobics program for total well-being: Exercise, diet, and emotional balance*. 3rd ed. Bantam; 1986.

25. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab*. 2012;10(2):486-489. https://doi.org/10.5812/ijem.3505

26. Poncet A, Courvoisier DS, Combescure C, Perneger TV. Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. *Methodol Eur J Res Methods Behav Soc Sci*. 2016;12(2):61-71. https://doi.org/10.1027/1614-2241/a000110

27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet Lond Engl*. 1986;1(8476):307-310.

28. Giavarina D. Understanding Bland Altman analysis. *Biochem Medica*. 2015;25(2):141-151. https://doi.org/10.11613/BM.2015.015

29. Hopkins WG. A Scale of Magnitudes for Effect Statistics. A New View of Statistics. August 7, 2006. Accessed August 16, 2018. http://sportsci.org/resource/stats/effectmag.html

30. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*. 2009;41(1):3-13. https://doi.org/10.1249/MSS.0b013e31818cb278

31. Buchheit M. Repeated-sprint performance in team sport players: associations with measures of aerobic fitness, metabolic control and locomotor function. *Int J Sports Med*. 2012;33(3):230-239. https://doi.org/10.1055/s-0031-1291364

32. Gabbett TJ, Jenkins DG, Abernethy B. Relative importance of physiological, anthropometric, and skill qualities to team selection in professional rugby league. *J Sports Sci*. 2011;29(13):1453-1461. https://doi.org/10.1080/02640414.2011.603348

33. Grant S, Corbett K, Amjad AM, Wilson J, Aitchison T. A comparison of methods of predicting maximum oxygen uptake. *Br J Sports Med*. 1995;29(3):147-152. https://doi.org/10.1136/bjsm.29.3.147

34. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: A systematic review. *PLoS ONE*. 2012;7(5):1-7. https://doi.org/10.1371/journal.pone.0037908

35. Lourenço TF, Silva FOC, Tessutti LS, Silva CE, Abad CCC. Prediction of 3000-m Running Performance Using Classic Physiological Respiratory Responses. *Int J Kinesiol Sports Sci*. 2018;6(3):18-24. https://doi.org/10.7575/aiac.ijkss.v.6n.3p.18

36. Sartor F, Bonato M, Papini G, et al. A 45-Second Self-Test for Cardiorespiratory Fitness: Heart Rate-Based Estimation in Healthy Individuals. *PloS One*. 2016;11(12):e0168154. https://doi.org/10.1371/journal.pone.0168154

37. Flouris AD, Metsios GS, Koutedakis Y. Enhancing the efficacy of the 20 m multistage shuttle run test. *Br J Sports Med*. 2005;39(3):166-170. https://doi.org/10.1136/bjsm.2004.012500

38. Chatterjee S, Chatterjee P, Bandyopadhyay A. Validity of Queen's College Step Test for estimation of maximum oxygen uptake in female students. *Indian J Med Res*. 2005;121(1):32-35.

39. Martínez-Lagunas V, Hartmann U. Validity of the Yo-Yo Intermittent Recovery Test Level 1 for direct measurement or indirect estimation of maximal oxygen uptake in female soccer players. *Int J Sports Physiol Perform*. 2014;9(5):825-831. https://doi.org/10.1123/ijspp.2013-0313

40. Metsios GS, Flouris AD, Koutedakis Y, Theodorakis Y. The effect of performance feedback on cardiorespiratory fitness field tests. *J Sci Med Sport*. 2006;9(3):263-266. https://doi.org/10.1016/j.jsams.2006.03.027